Systema Naturae, 2003, Vol. 5, pp. 219-236

UNA NUOVA DESCRIZIONE MATEMATICA DEL CODICE GENETICO

Diego L. Gonzalez* e Marcello Zanna#
**Fondazione Scuola di San Giorgio e Consiglio Nazionale delle Ricerche*
Isola di San Giorgio Maggiore, I-30124, Venezia, Italia.
e-mail diego.gonzalez@cini.ve.cnr.it
#ASL Bologna Sud, I-40053, Bologna, Italia.

Sommario

Dal punto di vista matematico, il codice genetico può essere visto come una mappa o applicazione suriettiva e non-iniettiva (vedasi Appendice 1 per le definizioni elementari della teoria degli insiemi qui utilizzate) tra l'insieme dei 64 possibili codoni (composti da tre basi) e l'insieme dei 21 elementi necessari alla sintesi delle proteine (i 20 aminoacidi più il segnale di Stop) (1). Come conseguenza di ciò, il codice è ridondante e degenerato. In analogia con il codice genetico, le rappresentazioni dei numeri interi del tipo "non-potenza" (non-power) (2) sono anch'esse delle mappe suriettive e non-iniettive tra insiemi di differente cardinalità e pertanto ridondanti. Ciò nonostante, nessuna delle rappresentazioni studiate finora descrive la degenerazione reale riscontrata nel codice genetico (3). In questo articolo viene descritto un nuovo tipo di rappresentazione numerica che porta alle seguenti sorprendenti conclusioni: (i) la degenerazione del codice genetico si può descrivere matematicamente, (ii) all'interno di questa degenerazione può essere riscontrata una nuova simmetria, (iii) assegnando a ogni codone un'appropriata stringa binaria, i codoni possono essere suddivisi in classi di parità definita (determinata anche dalla sequenza di basi del codone stesso). Quest'ultima proprietà è particolarmente suggestiva perché la codificazione di parità costituisce una delle strategie più semplici per la correzione degli errori nei sistemi elettronici di trasmissione di dati digitali (4).

La codificazione degli aminoacidi obbedisce a uno schema di tipo digitale: quattro simboli diversi raggruppati in gruppi di tre elementi, o triplette, codificano i 20 aminoacidi più il segnale di Stop, che indica la fine del processo di sintesi (1). Tale codifica non è altro che il codice genetico (vedasi Figura 1 per la versione standard del codice). Questa natura digitale della codificazione degli aminoacidi ha consentito lo sviluppo di modelli computazionali del DNA che hanno permesso, a loro volta, l'applicazione di tecniche della teoria dell'informazione (5) per l'analisi e l'interpretazione di lunghe sequenze di basi all'interno dei genomi.

	T	C	A	G	
T	TTT Phe	TCT Ser	TAT Tyr	TGT Cys	T
	TTC Phe	TCC Ser	TAC Tyr	TGC Cys	C
	TTA Leu	TCA Ser	TAA Stop	TGA Stop	A
	TTG Leu	TCG Ser	TAG Stop	TGG Trp	G
C	CTT Leu	CCT Pro	CAT His	CGT Arg	T
	CTC Leu	CCC Pro	CAT His	CGC Arg	C
	CTA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CTG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	ATT Ile	ACT Thr	AAT Asn	AGT Ser	T
	ATC Ile	ACC Thr	AAC Asn	AGC Ser	C
	ATA Ile	ACA Thr	AAA Lys	AGA Arg	A
	ATG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GTT Val	GCT Ala	GAT Asp	GGT Gly	T
	GTC Val	GCC Ala	GAC Asp	GGC Gly	C
	GTA Val	GCA Ala	GAA Glu	GGA Gly	A
	GTG Val	GCG Ala	GAG Glu	GGG Gly	G

Figura 1 Versione standard del codice genetico

Sebbene sistemi di numerazione quaternari (in base 4) siano stati utilizzati per rappresentare numericamente le triplette (6) (vedasi Appendice 2 per un esempio delle rappresentazioni dei numeri interi in una base arbitraria), dal punto di vista computazionale le codificazioni più efficienti sono del tipo binario (in base 2). Per rappresentare gli interi si utilizzano solo due diversi simboli (0,1) invece dei quattro necessari per rappresentare le quattro

basi della doppia elica di DNA (Timina, Citosina, Adenina e Guanina, o in modo abbreviato T,C,A,G). Per questo motivo fondamentale (ma non esclusivamente) diversi metodi sono stati proposti per convertire sequenze di basi in sequenze di bits (digiti binari). Quasi tutti utilizzano qualche forma di rappresentazione binaria fissa di due bits per ogni base all'interno della tripletta (in alcuni casi solo un bit quando si vuole evidenziare il carattere contrapposto di purina o pirimidina della base rappresentata) (7-11). Ciò nonostante, la descrizione binaria usuale rappresenta i numeri in maniera univoca e, a causa di questa implicita proprietà iniettiva (uno a uno), non è in grado di evidenziare l'eventuale ordine matematico nascosto in relazione con la degenerazione, e pertanto con il carattere suriettivo e non-iniettivo, del codice genetico (vedasi Appendice 1).

In questo senso, risultano molto più interessanti le cosiddette rappresentazioni binarie del tipo "non-potenza". In queste rappresentazioni, un dato numero può essere codificato simultaneamente da diverse stringhe binarie, vale a dire che esiste una degenerazione essenziale nella rappresentazione. A questo punto, pertanto, può essere formulata una domanda di tipo generale: *esiste una rappresentazione binaria del tipo "non-potenza" che descriva in modo completo la degenerazione del codice genetico?*. Per rispondere, bisogna fare ricorso alle proprietà generali di queste rappresentazioni e più in particolare alla proprietà di palindromia (vedasi Appendice 2).

La proprietà di simmetria palindromica assicura che, in una data rappresentazione non-potenza, esiste al massimo un unico sottoinsieme di numeri rappresentati con la stessa degenerazione il cui numero cardinale sia dispari. Ne consegue che la degenerazione attuale del codice genetico non può essere descritta in questo modo, perché esistono tre sottoinsiemi di aminoacidi che condividono la stessa degenerazione e il cui numero cardinale è dispari: 3 aminoacidi hanno degenerazione 6, 5 hanno degenerazione 4, e 9 hanno degenerazione 2 (vedasi Tabella 1a). A questo punto, però, è necessario notare che ciò che sembra essere determinante non è la degenerazione totale, ma la degenerazione all'interno dei quartetti, cioè all'interno dei gruppi di quattro triplette determinate dalle loro due prime lettere (11-13). Questo equivale a dire che gli aminoacidi di degenerazione 6 sono in realtà codificati da due sottoinsiemi: uno con degenerazione 2 e un altro con degenerazione 4. In tal caso si può dimostrare che la rappresentazione binaria "non-potenza" del codice genetico è possibile se si assegnano alle 6 basi numeriche posizionali i valori $(x_1, x_2, \ldots x_6) = (1,1,2,4,7,8)$.

Degenerazione	# di aminoacidi
6	3
4	5
3	2
2	9
1	2

Tabella 1a

Degenerazione	# di aminoacidi
4	8
3	2
2	12
1	2

Tabella 1b

Tabella 1a - Numero di aminoacidi che condividono la stessa degenerazione (stesso numero di triplette che le codificano).

Tabella 1b - Numero di aminoacidi che condividono la stessa degenerazione all'interno dei quartetti, cioè, dei gruppi di quattro triplette che condividono le stesse due lettere iniziali, ad esempio, TCT, TCC, TCA e TCG.

Questa rappresentazione particolare si può chiamare "*rappresentazione binaria non-potenza tipo DNA*" (vedasi Figura 2). La rappresentazione tipo DNA descrive tutte le degenerazioni riscontrate all'interno dei quartetti nel codice genetico standard (vedasi Tabella 1b): ci sono 2 oggetti con degenerazione 1, 12 con degenerazione 2, 2 con degenerazione 3 e 8 con degenerazione 4 (gli aminoacidi con degenerazione 6 contribuiscono con 3 elementi alle famiglie di degenerazione 4 e 2).

Calcoli fatti su altre corrispondenze simili (dal punto di vista delle trasformazioni matematiche coinvolte nella rappresentazione) dimostrano che la probabilità che la coincidenza riportata sia dovuta al caso è molto bassa (14). Ciò nonostante, senza un collegamento tra le stringhe binarie della rappresentazione tipo DNA e le triplette reali, sembra difficile poter cogliere il significato biologico di questo ordine nascosto. Un primo tentativo per stabilire questa corrispondenza tra le stringhe binarie e le triplette del codice genetico è basato sulle proprietà di simmetria di entrambi, ed è anch'esso rappresentato nella Figura 2. I principali argomenti che stanno alla base dell'assegnazione riportata possono essere riassunti brevemente in due punti:

(1) *Codificazione della terza lettera della tripletta*: tutte le versioni conosciute del codice genetico possiedono una simmetria perfetta nello scambio delle basi T e C (C/T) nell'ultima lettera della tripletta. Nella Figura 2 possiamo osservare che lo stesso è valido per le stringhe che finiscono in 0,1 o 1,0;

Numero rappresentato	Stringhe binarie																								Aminoacido
	8	7	4	2	1	1	8	7	4	2	1	1	8	7	4	2	1	1	8	7	4	2	1	1	
0	0	0	0	0	0	0																			W Trp
1	0	0	0	0	1	0	0	0	0	0	0	1													F Phe
2	0	0	0	0	1	1	0	0	0	1	0	0													Ter
3	0	0	0	1	0	1	0	0	0	1	1	0													Y Tyr
4	0	0	1	0	0	0	0	0	0	1	1	1													L Leu 2
5	0	0	1	0	0	1	0	0	1	0	1	0													H His
6	0	0	1	1	0	0	0	0	1	0	1	1													Q Gln
7	0	1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	0	1							C Cys
8	0	0	1	1	1	1	0	1	0	0	0	1	0	1	0	0	1	0	1	0	0	0	0	0	S Ser 4
9	1	0	0	0	0	1	1	0	0	0	1	0	0	1	0	1	0	0	0	1	0	0	1	1	P Pro
10	1	0	0	1	0	0	0	1	0	1	0	1	0	1	0	1	1	0	1	0	0	0	1	1	V V al
11	0	1	1	0	0	0	0	1	0	1	1	1	1	0	0	1	0	1	1	0	0	1	1	0	L Leu 4
12	1	0	1	0	0	0	1	0	0	1	1	1	0	1	1	0	1	0	0	1	1	0	0	1	R Arg 4
13	0	1	1	0	1	1	1	0	1	0	0	1	1	0	1	0	1	0	0	1	1	1	0	0	G Gly
14	0	1	1	1	1	0	0	1	1	1	0	1	1	0	1	0	1	1	1	0	1	1	0	0	A Ala
15	1	1	0	0	0	0	1	0	1	1	0	1	1	0	1	1	1	0	0	1	1	1	1	1	T Thr
16	1	0	1	1	1	1	1	1	0	0	0	1	1	1	0	0	1	0							I Ile
17	1	1	0	0	1	1	1	1	0	1	0	0													E Glu
18	1	1	0	1	0	1	1	1	0	1	1	0													D Asp
19	1	1	0	1	1	1	1	1	1	0	0	0													R Arg 2
20	1	1	1	0	0	1	1	1	1	0	1	0													N Asn
21	1	1	1	1	0	0	1	1	1	0	1	1													K Lys
22	1	1	1	1	0	1	1	1	1	1	1	0													S Ser 2
23	1	1	1	1	1	1																			M Met

Figura 2 - Rappresentazione binaria non-potenza tipo DNA e codice genetico:

Rappresentazione dei primi 24 numeri interi (mostrati nella colonna a sinistra) nella rappresentazione non-potenza tipo DNA. La degenerazione del codice genetico all'interno dei quartetti è descritta in modo esatto dalla rappresentazione numerica; per un dato numero intero la degenerazione corrisponde al numero di stringhe binarie di lunghezza 6 (compilate orizzontalmente). Nella colonna di destra sono rappresentati gli aminoacidi assegnati in base a considerazioni di simmetria (versione euplotidea nucleare del codice genetico, Figura 3). Una particolare simmetria che conserva la degenerazione, la simmetria palindromica, risulta evidente come una simmetria speculare tra i due gruppi di stringhe (come se uno specchio immaginario fosse collocato orizzontalmente tra la parte superiore e inferiore della figura). A livello aritmetico il palindromo di una stringa è ottenuto mediante l'operazione di complemento a 1, vale a dire scambiando 1/0 per 0/1. A causa di questa perfetta simmetria la posizione assoluta di un dato aminoacido nella tabella non è completamente determinata. I singoli aminoacidi possono essere scambiati all'interno di una coppia palindromica (i due aminoacidi assegnati al numero R, e al (23-R)) senza modificare la degenerazione. La parità delle stringhe è anch'essa rappresentata nella figura: stringhe pari sono colorate in rosso e stringhe dispari sono colorate in verde (corrispondenti a un numero pari o dispari di 1 nelle stringhe; vedasi Appendice 2).

tutti i numeri codificati da queste stringhe rimangono invariati dalla trasformazione 0,1/1,0. Pertanto possiamo concludere che lo scambio a_6,a_5,a_4,a_3,0,1/a_6,a_5,a_4,a_3,1,0, a livello delle stringhe binarie, è equivalente allo scambio C/T nell'ultima lettera della tripletta; il numero codificato dalle rispettive stringhe o l'aminoacido codificato dalle triplette rimangono invariati in queste trasformazioni.

Per esclusione, si può concludere che le stringhe che finiscono in 0,0 o 1,1 devono codificare triplette che finiscono in A o G. Ma osservando che i due aminoacidi di degenerazione 1 (Triptofano e Metionina) sono codificati necessariamente dalle stringhe (0,0,0,0,0,0) e (1,1,1,1,1,1), e che entrambi hanno delle triplette che finiscono in G, si può dedurre che non basta la terminazione in 0,0 o 1,1 per definire la lettera finale associata a questo tipo di triplette (che finiscono in A o G); diventa necessario prendere in considerazione la parità delle stringhe. Infatti, *stringhe pari che finiscono in 0,0 o 1,1 sono associate a triplette che finiscono in G e stringhe dispari che finiscono in 0,0 o 1,1 ad una A finale*. Questa assegnazione rappresenta la prima regola di parità. Ricordiamo che la parità di una stringa binaria si definisce in base alla quantità di 1 contenuti nella stessa: un numero pari di 1 definisce una stringa pari, un numero dispari di 1, una stringa dispari (vedasi Appendice 2).

(2) *Simmetria palindromica*: applicando le regole suddette e sapendo che i due aminoacidi di degenerazione 3 sono necessariamente rappresentati dalle due stringhe di degenerazione 3, si arriva alla conclusione che le triplette corrispondenti dovrebbero finire in T, C e A. Questo è vero per le triplette che codificano per l'aminoacido I (Ile) però non lo è per il segnale di Stop. Il segnale di Stop è l'unico elemento con degenerazione minore o eguale a quattro che presenta triplette con diverse seconde lettere. Infatti, la tripletta TGA presenta delle caratteristiche anomale. Ad esempio, può codificare diversi aminoacidi in diverse versioni del codice e, in alcuni casi, un 21-esimo aminoacido, la Selenocisteina (15-16). Pertanto si può pensare che quest'assegnazione nel codice standard rappresenti una specie di rottura di simmetria, com'è stato suggerito in diversi contesti (17-18). Possiamo osservare anche che, in una versione particolare del codice genetico, quella corrispondente al genere *Euplotes* (nucleare), la tripletta TGA rappresenta l'aminoacido C (Cys), il quale risulta allora uno dei due aminoacidi di degenerazione 3 (invece del segnale di Stop) rappresentato da tre triplette che finiscono in T, C e A, come risulta implicito nella rappresentazione numerica tipo DNA (Figura 3).

	T	C	A	G	
T	TTT Phe	TCT Ser	TAT Tyr	TGT Cys	T
	TTC Phe	TCC Ser	TAC Tyr	TGC Cys	C
	TTA Leu	TCA Ser	TAA Stop	TGA Cys	A
	TTG Leu	TCG Ser	TAG Stop	TGG Trp	G
C	CTT Leu	CCT Pro	CAT His	CGT Arg	T
	CTC Leu	CCC Pro	CAT His	CGC Arg	C
	CTA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CTG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	ATT Ile	ACT Thr	AAT Asn	AGT Ser	T
	ATC Ile	ACC Thr	AAC Asn	AGC Ser	C
	ATA Ile	ACA Thr	AAA Lys	AGA Arg	A
	ATG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GTT Val	GCT Ala	GAT Asp	GGT Gly	T
	GTC Val	GCC Ala	GAC Asp	GGC Gly	C
	GTA Val	GCA Ala	GAA Glu	GGA Gly	A
	GTG Val	GCG Ala	GAG Glu	GGG Gly	G

Figura 3 - Versione del codice genetico corrispondente al genere Euplotes (nucleare). Il genere Euplotes appartiene alla classe Nassophorea nel phylum Ciliophora (cigliati). In grigio si evidenza l'unica tripletta che differisce nell'assegnazione rispetto al codice genetico standard, la TGA, che viene assegnata all'aminoacido Cisteina invece che al segnale di Stop.

Per questo motivo, possiamo considerare nel presente approccio la versione corrispondente al genere *Euplotes* del codice, come una struttura generale sulla quale differenti rotture di simmetria possono descrivere le altre versioni, e in particolare, la versione standard che differisce dalla *Euplotes* solo nell'assegnazione della tripletta TGA (Figura 3). Possiamo osservare che, nel caso della versione *Euplotes* del codice genetico, la simmetria palindromica è rappresentata da trasformazioni fra quartetti (gruppi di quattro triplette) che preservano la degenerazione (Figura 4). In un certo senso, le trasformazioni palindromiche sono complementari delle trasformazioni di Rumer (11) che collegano quartetti di differenti degenerazioni.

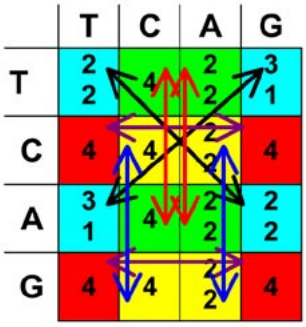

Figura 4 - Simmetria palindromica. Tutti i quartetti (definiti da 2 identiche prime lettere nella tripletta) sono collegati a coppie dalla trasformazione palindromica). Frecce dello stesso colore indicano un'operazione comune a livello delle triplette. L'insieme completo delle trasformazioni può essere scritto in forma compatta come segue:

	prima lettera	seconda lettera	terza lettera
trasformazione	[T↔A ; Č ; Ğ]	[T↔G]	[T↔C; Å; Ğ]
trasformazione	[T↔A ; C↔G]	[Å ; Č]	[T↔C; Å; Ğ]

Queste trasformazioni non sono state riportate prima (le cosiddette trasformazioni di Rumer (11) sono simili ma non preservano la degenerazione dei quartetti collegati dall'operazione di simmetria). Si può osservare che la lettera T (Timina) è l'unica che viene trasformata in tutte le posizioni all'interno del codone. La simmetria corrisponde esattamente alla versione Euplotes del codice genetico. Il codice standard può essere visto come una rottura di simmetria a livello del codone TGA, che in tal caso codifica per il segnale di STOP invece di Cisteina.

Il fatto notevole nel presente approccio è che le trasformazioni palindromiche vengono descritte a livello matematico da una regola semplice e precisa: due triplette sono palindromiche se sono rappresentate da stringhe binarie collegate dalla operazione di complemento a 1 (rimpiazzare ogni 1/0 con 0/1). L'insieme completo delle trasformazioni palindromiche per la versione "euplotidea" del codice, è rappresentata graficamente per le due prime lettere delle triplette nella Figura 4 (si sottolinea che il termine simmetria palindromica si riferisce qui alle proprietà matematiche delle stringhe binarie che rappresentano le triplette e non alla sequenza reale di basi lungo la catena di DNA).

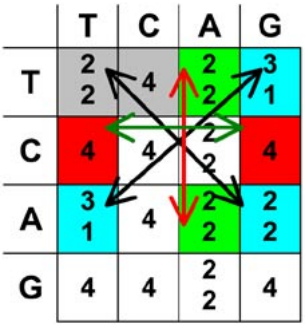

Figura 5 - Varianti del codice genetico e trasformazione palindromica.
Le diverse versioni conosciute del codice genetico, sono evidenziate in colore. Vi sono 8 quartetti variabili e tutti sono relazionati tra di loro per qualche trasformazione palindromica. I quartetti coinvolti in qualche variante degli aminoacidi codificati o del segnale di STOP sono evidenziati in colore. Le uniche eccezioni sono evidenziate in grigio: TCN, per la quale è riportata una variazione nel codone TCA (20), e TTN per la quale non sono state riportate delle variazioni. In modo analogo, i quartetti invarianti sono anch'essi collegati dalle trasformazioni palindromiche.

Poiché le trasformazioni palindromiche non cambiano la parità, il risultato per la terza lettera delle triplette è triviale: C è scambiata con T mentre A e G rimangono invariate.

Molte proprietà interessanti della rappresentazione tipo DNA saranno discusse in un lavoro più esteso (D.L. Gonzalez, in preparazione). Qui possiamo sottolineare solo altri due fatti relazionati con la simmetria palindromica:

(a) *Posizione delle triplette variabili nelle differenti versioni del codice genetico*. Nella Figura 5 abbiamo rappresentato le variazioni nell'assegnazione degli aminoacidi alle diverse triplette per tutte le versioni conosciute del codice genetico (19) prendendo come struttura di riferimento quella corrispondente alla versione Euplotes (si può sottolineare che la posizione dei siti variabili rappresenta un sistema di riferimento assoluto). È notevole che tutte le variazioni d'assegnazione accadono in quartetti collegati dalla trasformazione palindromica che definisce esattamente due metà del codice genetico composte da 8 quartetti: una metà è proclive alla riassegnazione delle triplette e l'altra no. Solo una delle 24 varianti riportate nel referimento 18 sfugge a questa regola (20).

(b) *Regole di parità*. L'insieme delle trasformazioni palindromiche è compatibile con la seguente regola di parità per le triplette che finiscono in T o C: le triplette che finiscono in T o C e che hanno una parità totale dispari, hanno una T o una G come seconda lettera; quelle con parità totale pari, hanno una C o una A. Questa regola è complementare a quella precedente per triplette che finiscono in A o G e mostra che tutte le triplette sono marcate riguardo alla loro parità: se finiscono in G o in A, il segno è coincidente con detta lettera; se invece finiscono in T o in C, la parità è codificata nella seconda lettera della tripletta (vedasi Figura 6). Di nuovo emerge un fatto notevole, e cioè, che una proprietà matematica semplice e concisa, in questo caso la parità delle stringhe, è in relazione con proprietà definite delle triplette in termini della organizzazione dei nucleotidi lungo la doppia elica del DNA. Sembrerebbe che le triplette seguano un'organizzazione molto strutturata, che si basa su semplici regole matematiche a livello delle stringhe binarie che le rappresentano.

In questo lavoro si è dimostrato che un particolare sistema di rappresentazione dei numeri interi, cioè la rappresentazione non-potenza tipo DNA, descrive in modo esatto le degenerazioni riscontrate nel codice genetico. Si è visto, inoltre, che questo sistema consente una rappresentazione binaria delle triplette. Tale codificazione binaria non è fissa ma dipende dal contesto: la codificazione binaria di una base specifica dipende dalle altre lettere nella tripletta e anche dalle posizioni relative al suo interno. Questo fatto, unitamente alla codificazione di parità per mezzo della seconda o terza lettera della tripletta, suggerisce fortemente l'esistenza di un meccanismo di correzione degli errori basato sul controllo della parità. Questa possibilità è stata suggerita (21) e, almeno da un punto di vista lineare, investigata (6). E' da sottolineare che una strategia di codificazione di parità è stata recentemente suggerita per rendere conto della selezione di basi complementari nella doppia elica di DNA (22).

Il nostro approccio suggerisce che anche "lungo" la doppia elica esista un meccanismo di correzione degli errori e che questo meccanismo contribuisca all'accuratezza della sintesi delle proteine. Infatti, a differenza della codificazione di parità delle coppie complementari che hanno un significato solo in un contesto evoluzionistico, il meccanismo proposto può risultare attivo nel meccanismo di trascrizione/sintesi e, pertanto, può essere responsabile del bassissimo tasso di errori nella trascrizione e dei diversi tassi di errori a seconda della posizione dei nucleotidi nel sistema di riferimento per la lettura. Questo fatto può avere un particolare

	T	C	A	G	
T	TTT Phe	TCT Ser	TAT Tyr	TGT Cys	T
	TTC Phe	TCC Ser	TAC Tyr	TGC Cys	C
	TTA Leu	TCA Ser	TAA Stop	TGA Cys	A
	TTG Leu	TCG Ser	TAG Stop	TGG Trp	G
C	CTT Leu	CCT Pro	CAT His	CGT Arg	T
	CTC Leu	CCC Pro	CAT His	CGC Arg	C
	CTA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CTG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	ATT Ile	ACT Thr	AAT Asn	AGT Ser	T
	ATC Ile	ACC Thr	AAC Asn	AGC Ser	C
	ATA Ile	ACA Thr	AAA Lys	AGA Arg	A
	ATG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GTT Val	GCT Ala	GAT Asp	GGT Gly	T
	GTC Val	GCC Ala	GAC Asp	GGC Gly	C
	GTA Val	GCA Ala	GAA Glu	GGA Gly	A
	GTG Val	GCG Ala	GAG Glu	GGG Gly	G

Figura 6 - Segnalazione di parità nelle triplette. Utilizzando lo stesso codice di colore della Figura 2, si evidenziano triplette pari e dispari. Quelle che finiscono in A o G sono rispettivamente dispari e pari; le triplette che finiscono in C o T invece, hanno la parità determinata dalla seconda lettera: T o G determinano triplette dispari, mentre C o A determinano triplette pari. Si può osservare che c'è una perfetta corrispondenza con la parità delle stringhe binarie riportate nella Figura 2.

significato in medicina, dato che un'anomalia in tale sistema di correzione può produrre un incremento nel numero degli errori o l'impossibilità di correggere errori specifici associati a specifiche malattie. Inoltre, una semplice proteina con una lunghezza di 100 aminoacidi, può essere codificata in 3^{100} modi, cioè approssimativamente in 10^{50} modi diversi!

Come e perché una sequenza particolare di nucleotidi sia scelta per codificare una specifica proteina, rimane una delle domande più importanti sul problema dell'organizzazione dell'informazione nel DNA. Il presente approccio offre un'opportunità per esplorare a fondo alcuni di questi aspetti organizzativi della codificazione dell'informazione biologica.

Il suggerimento più naturale, in questo senso, è quello di analizzare le proprietà statistiche delle sequenze reali di triplette codificanti il DNA assegnando loro le stringhe binarie della rappresentazione tipo DNA prima descritta. Possono essere studiate anche altre statistiche semplificate, ad esempio utilizzando solo le proprietà di parità delle triplette (un unico bit per tripletta).

Dal punto di vista teorico, sono state proposte diverse teorie per spiegare le regolarità riscontrate nel codice genetico (23). In alcuni casi queste possono essere associate alle proprietà fisico-chimiche delle molecole coinvolte nella codifica e decodifica degli aminoacidi (11, 16, 23). Nonostante ciò, rimane sconcertante il fatto che diverse proprietà fondamentali del codice genetico, come la distribuzione della degenerazione, e anche insospettate proprietà nascoste, come la simmetria palindromica o la parità delle triplette, riflettano un ordine matematico profondo che viene descritto esattamente da una delle operazioni più elementari e semplici che stanno alle radici della matematica: la rappresentazione dei numeri.

BIBLIOGRAFIA

1. Brown, T.A., *Genome*, Second Edition, BIOS Scientific Publishers, Oxford, (2002).
2. Wolfram, S., *A New Kind of Science,* Wolfram Media, Illinois, (2002).
3. Knuth, D.E., *The Art of Computer Programming,* volume 2, *Seminumerical Algorithms*, third edition, Addison Wesley, Reading, Massachusetts (1997).
4. Sweeney, P., *Error Control Coding*, Wiley, New York, (2002)
5. Yockey, H.P., *Information Theory and Molecular Biology*, Cambridge University Press, Cambridge, New York, (1992).
6. Liebovitch L.S., Tao Yi, Todorov A.T. and Levine L., Is there an error-correcting code in the base sequence in DNA?, *Biophysical Journal*, **71**, 1539-1544 (1996).
7. Jimenez-Montaño, M.A., de la Mora-Basañez, C.R. and Poschel, T., The hypercube structure of the genetic code explains conservative and non-conservative aminoacid substitutions in vivo and in vitro, *Biosystems*, **39**, 117- 125 (1996).
8. Jimenez-Montaño, M.A., de la Mora-Basañez, C.R., The genetic code as a six-dimensional boolean hypercube. In: *Abstracts of Proc. Soc. Math. Biol. Annual Meeting*, July 23-26, U.C., Berkeley, CA (1992).

9. Karasev, V.A. and Sorokin, S.G., Topological structure of the genetic code, *Russ. J. Genet.*, **33**, 622-628 (1997).
10. Klump, H.H., The physical bases of the genetic code: the choice between speed and precision, *Archv. Biochem. Biophys.*, **301**, 207-209 (1993).
11. Rumer, Y.B. About the codon's systematization in the genetic code (in Russian), *Proc. Acad. Sci. U.S.S.R.* (Doklady) **167**, 1393 (1966).
12. Lehmann, J., Physico-chemical constraints connected with the coding properties of the genetic system*, J. Theor. Biol.*, **202**, 129-144 (2002).
13. Shcherbak, V.I., The symmetrical architecture of the genetic code systematization principle, *J. Theor. Biol.*, **162**, 395-398 (1993).
14. Un calcolo riguardante la possibilità di generazione casuale delle cosidette *trasformazioni di Rumer* (le quali presentano alcune analogie con il presente approccio, nel senso che le trasformazioni palindromiche mettono anch'esse in relazione gruppi di 8 quartetti), danno una probabilità minore di 3.09^{-32}; Zhaxybayeva, O., Statistical estimation of Rumer's transformation of the universal genetic code, *ISSOL'96*, Orleans, France (1996).
15. Zinoni, F., Birkmann, A., Stadtman, T.C., and Bock, A, Nucleotide sequence and expression of the selenocysteine-containing polypeptide of formate dehydrogenase (formate-hydrogen-lyase-linked) from *Escherichia coli, Proc. Nat. Acad. Sci.*, **83**, 4650 (1986).
16. Atkins, J.F. and Gesteland, R.F., Selenocysteine, the 21st amino acid, *Nature*, **407**, 463-465 (2000).
17. Jimenez-Montaño M.A., Protein evolution drives the evolution of the genetic code and vice versa, *Biosystems*, **54**, 47-64 (1999).
18. Hornos, J.E.M. and Hornos Y.M.M., Algebraic model for the evolution of the genetic code, *Phys. Rev. Lett.*, **71**, 4401-4404 (1993).
19. Knight, R.D., Freeland, S.J and Landweber, L.F., Rewiring the Keyboard: Evolvability of the Genetic Code, *Nature Reviews – Genetics*, **2**, 49-58 (2001).
20. Kuck, U., Jekosch, K. and Holtzamer, P., DNA sequence analysis of the complete mitochondrial genome of the green alga *Scnenedesmus obliquus*: evidence for UAG being a leucine and UCA being a non-sense codon, *Gene*, **253**, 13-18, (2000).
21. Forsdyke, D., Are introns in-series error-detecting sequences*, J. Theor. Biol.*, **93**, 861-866 (1981).
22. Donall A. Mac Donaill, A parity code interpretation of nucleotide alphabet composition, *Chem. Commun.*, **18**, 2062-2063 (2002).

23. Di Giulio, M., On the origin of the genetic code, *J. Theor. Biol.*, **187**, 573-581 (1997).
24. O. Ore, *Number Theory and its History*, Dover Publications, New York (1988).

APPENDICE 1
Proprietà delle applicazioni tra insiemi

In genere, una operazione logica che mette in relazione, o "applica", gli elementi di un insieme A con quelli di un altro insieme B, viene chiamata "applicazione" (in inglese *mapping*). A livello grafico la si può rappresentare mediante frecce che collegano gli elementi dei due insiemi. In base ad alcune proprietà generali, le applicazioni possono essere classificate come:

(i) *iniettive*: se ogni elemento dell'insieme B proviene da uno o da nessun elemento di A (vedasi Figura 7);

(ii) *suriettive*: se ogni elemento di B proviene da uno o da più elementi di A (in B non ci sono elementi vacanti, vedasi Figura 8);

(iii) *biiettive*: se valgono simultaneamente le due proprietà suddette.

Dal punto di vista della teoria degli insiemi, sia il codice genetico che le rappresentazioni dei numeri del tipo "non-potenza" sono, allo stesso tempo, *suriettive* (tutti gli elementi dell'insieme B, aminoacidi o numeri interi, provengono da almeno un elemento dell'insieme A, cioè codoni o stringhe binarie) e *non-iniettive* (alcuni elementi dell'insieme B provengono da più di un elemento dell'insieme A, quindi ridondanza e degenerazione sono implicite, Figura 8).

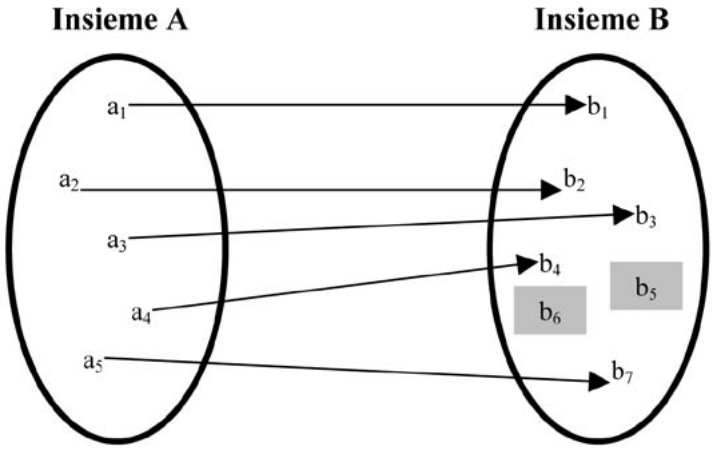

Figura 7 - Applicazione iniettiva. Nessun elemento di B "proviene" da più di un elemento di A, ma ci possono essere elementi di B vacanti (indicati in grigio).

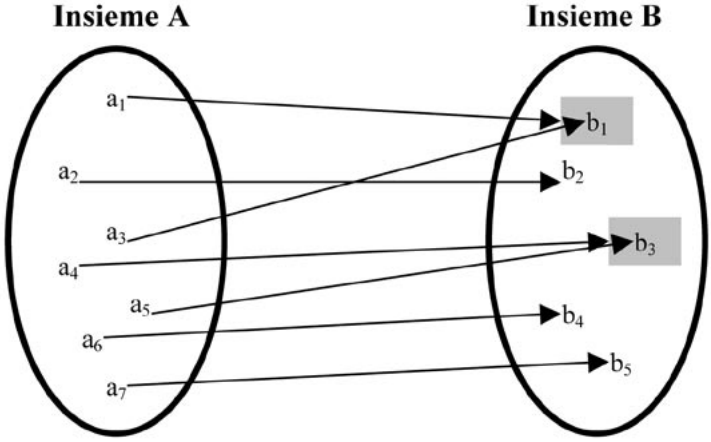

Figura 8 - Applicazione suriettiva. Nessun elemento di B è "vacante". Se alcuni elementi di B provengono da più di un elemento di A (indicati in grigio) l'applicazione è anche non-iniettiva (nell'esempio, due elementi di B hanno degenerazione 2).

APPENDICE 2
Rappresentazione dei numeri interi

Le rappresentazioni degli interi utilizzano in genere le potenze di una base k, pesate con dei digiti che coprono il range (0; k-1)[24]. Questo assicura che la rappresentazione è univoca (uno a uno). Nella rappresentazione digitale viene utilizzata la base 10 (k=10) e i coefficienti posizionali, che possono prendere i valori tra 0 e 9, sono collocati seguendo l'ordine ascendente delle potenze di 10. Ad esempio, la rappresentazione in base 10 del numero 15 viene rappresentata come segue:

n		$10^1 = 10$	$10^0 = 1$
15		1	5

$15 = 1.10 + 5.1 = 10 + 5 = 15$

La rappresentazione binaria[24] corrisponde alla base k=2, pertanto i valori posizionali corrispondono alle potenze di due e i relativi pesi o coefficienti possono assumere solo i valori 0 o 1. Il numero 15 viene rappresentato nella base binaria nel modo seguente:

n		$2^3 = 8$	$2^2 = 4$	$2^1 = 2$	$2^0 = 1$
15		1	1	1	1

$15 = 1.8 + 1.4 + 1.2 + 1.1 = 8 + 4 + 2 + 1 = 15$

Una generalizzazione dei sistemi basati sulle potenze di una base k consiste nell'assegnare valori arbitrari a questi valori posizionali (invece delle potenze di k). Per questo motivo tali sistemi possono essere chiamati "non-potenza" (non-power)[2]. Se questi valori posizionali arbitrari crescono più lentamente delle potenze di 2, la rappresentazione è in generale completa (tutti i numeri da 0 al numero formato dalla somma di tutti i valori posizionali, sono rappresentati) ma ridondante (un dato numero può essere rappresentato da più di una stringa binaria). Un esempio interessante di questo tipo lo costituisce la cosiddetta rappresentazione di Fibonacci[2]. In questa rappresentazione, i valori delle basi posizionali corrispondono

a successivi numeri di Fibonacci. I numeri di Fibonacci prendono il nome dal suo scopritore, Leonardo Pisano, il famoso matematico di Pisa noto anche come Fibonacci. I numeri di Fibonacci formano una serie nella quale l'ennesimo numero è ottenuto come somma dei due precedenti, e cioè, $F_n = F_{n-2} + F_{n-1}$, con la condizione iniziale $F_1 = 1$ and $F_2 = 1$. La rappresentazione di Fibonacci di ordine 6 (parole binarie di lunghezza 6 o 6 bits) utilizza i primi 6 numeri di Fibonacci, 1, 1, 2, 3, 5, 8. Il numero 15, ad esempio, è rappresentato nel modo seguente,

n		8	5	3	2	1	1		8	5	3	2	1	1		8	5	3	2	1	1
15		1	1	0	1	0	0		1	1	0	0	1	1		1	0	1	1	1	1

15 = 1.8+1.5+1.2 = 8 + 5 +2 = 15
15 = 1.8+1.5+1.1+1.1 = 8 + 5 + 1 + 1 = 15
15 = 1.8+1.3+1.2+1.1+1.1 = 8 + 3 + 2 + 1 + 1 = 15

Abbiamo sottolineato in colore il concetto di parità di una stringa: stringhe pari sono evidenziate in rosso e stringhe dispari in verde. La parità si definisce in base al numero di 1 in una data stringa: un numero pari da una stringa pari, un numero dispari da una stringa dispari. La rappresentazione di Fibonacci di ordine 6 presenta diverse proprietà in comune con il codice genetico. Ad esempio, è ridondante (il numero 15 è rappresentato da 3 diverse stringhe binarie) e vi sono esattamente 21 numeri rappresentati da 64 stringhe binarie (come i 20 aminoacidi più il segnale di STOP rappresentati dai 64 codoni). Ciò nonostante, la degenerazione della rappresentazione di Fibonacci di ordine 6 non coincide con la degenerazione osservata nel codice genetico (vedasi Figura 1 per una descrizione completa della degenerazione del codice).
In un contesto più largo, le rappresentazioni non-potenza hanno una proprietà che le caratterizza: la degenerazione è una funzione palindroma dei numeri rappresentati; i numeri r, and R-r, dove R è il massimo intero che può essere rappresentato, condividono la stessa degenerazione. Le coppie relazionate dalla simmetria palindromica sono rappresentate numericamente da stringhe complementari: il palindromo di una data stringa si ottiene scambiando simultaneamente tutti gli 1 per 0 e viceversa. Una conseguenza importante di questa proprietà è che il sottoinsieme di numeri che condividono la stessa degenerazione ha un numero cardinale

pari per tutte le degenerazioni della rappresentazione. Una sola eccezione a questa regola è possibile: nel caso la somma di tutte le basi posizionali sia un numero pari R, il numero centrale della rappresentazione, cioè, R/2, possiede una degenerazione che è condivisa da un numero dispari di numeri rappresentati.